Improving short-term, near-surface temperature forecasts by integrating weather pattern information into Model Output Statistics

Matthias Zech [a], Lueder von Bremen [a]

[a] *German Aerospace Center, Institute of Networked Energy Systems*

Corresponding author: Matthias Zech, matthias.zech@dlr.de

ABSTRACT: Dynamical numerical weather prediction has remarkably improved over the last decades. Yet, postprocessing techniques are needed to calibrate forecasts which are based on statistical and Machine Learning techniques. With recent advances in the derivation of year-round, large-scale atmospheric circulations, or weather regimes, the question arises of whether this information can be valuable within forecast postprocessing methods. This paper investigates this by proposing a bias correction scheme to integrate the atmospheric circulation state derived from empirical orthogonal functions, referred to as weather patterns, for deterministic short-term, near-surface temperature forecasts based on LASSO regression. We propose a computational study which first evaluates different weather pattern definitions (spatial domain) to improve temperature forecasts in Europe. As a bias could be associated with the weather pattern at the model initialization time or at the realization time of the forecast, both variants are tested in this study. We show that forecasted weather patterns with the identical spatial domain as the forecast show best skill reaching Mean Squared Error Skill improvements of up to 3% (day-ahead) or 1% respectively (week ahead). Only considering land surface improvements in Europe, improvements of 4-6% for day-ahead and 1 to 5% for week-ahead forecasts are observable. We believe that this study not only introduces a simple yet effective tool to reduce bias in temperature forecasts but also contributes to the active discussion of how valuable weather patterns are and how to use them within forecast calibration techniques.

1. Introduction

Dynamical numerical weather prediction (NWP) models have remarkably improved over the last decades with skill improvements of approximately one day per decade in the range of 3-day to 10-day forecasts meaning that a 4-days-ahead forecast is today approximately as accurate as a 3-days-ahead forecast ten years ago (Bauer et al. 2015). However, NWP forecasts have deficiencies originating from the difficulty in determining initial conditions, boundary condition errors, and model structural errors that increase with longer lead times (Vannitsem et al. 2021; Bauer et al. 2015). Furthermore, dynamical models suffer from bias and dispersion, which requires statistical post-processing techniques for NWP forecasts (Vannitsem et al. 2021). Statistical models have been successfully applied to calibrate forecasts, which are initially introduced by the Model Output Statistics (MOS) technique from Glahn and Lowry (1972) which combines dynamical and statistical models through linear regressions. With the rise of ensemble forecasts, the MOS technique has been extended by Gneiting et al. (2005) to the Ensemble Model Output Statistics (EMOS) method, which is a non-homogeneous regression applied in a rolling training window fashion (Gneiting et al. 2005). Both MOS and EMOS have shown remarkable success in improving the forecast skill of different meteorological parameters.

Recent studies show that including additional atmospheric variables in addition to the variable of interest can enhance the performance of post-processing techniques (Messner et al. 2017; Rasp and Lerch 2018a; Taillardat et al. 2016). However, choosing the appropriate variables is a difficult task, as more variables increase the model complexity and the risk of overfitting. Furthermore, with fewer exogenous variables the post-processing model interpretability increases and can give useful information about situations in which NWP models are more likely to be erroneous and require corrections. Allen et al. (2019) therefore propose to use weather regimes as a description of the large-scale atmospheric situation and to integrate this information into forecast post-processing techniques. Weather regimes are based on the premise that the chaotic nature of weather can be decomposed into a finite set of quasi-stationary, recurrent and persistent atmospheric flow patterns (Michelangeli et al. 1995; Grams et al. 2020). They have a long tradition in synoptic-scale dynamic meteorology first mentioned by Rex (1950, 1951) who analyze the impact of atmospheric blocking events on European precipitation and surface temperatures. Since then, weather regimes have succeeded in explaining the frequency and magnitude of temperature and precipitation events (Robertson and Ghil 1999), the probability of extreme events (Robertson and Ghil 1999), and the production of renewable energy (Grams et al. 2017; Van Der Wiel et al. 2019). Within synoptic-scale forecasting, several authors observe a dependency between NWP low-frequency forecast errors and the underlying weather regime (Koch 1985; O'Lenic and Livezey 1989; Stoss and Mullen 1995). Ferranti et al. (2015) show skill differences for short- and mid-term ensemble forecasts and observe the lowest forecast skill during blocking events raised by the lack of the NWP model to predict the transition from and into blocking events, including their persistence. The relevance of blocking events for the NWP forecast skill is already observed in Tibaldi and Molteni (1990) but is still relevant today, since forecast busts are often associated with blocking anticyclones in operational NWP models (Rodwell et al. 2013; Grams et al. 2018). The relationship between weather regimes and forecast accuracy is also observed for extended winter periods showing significant differences in NWP forecast errors during different flow patterns (Ferranti et al. 2015). This motivates Allen et al. (2019) to extend ensemble post-processing techniques, namely nonhomogeneous regressions (Gneiting et al. 2005) and Bayesian Model Averaging (Raftery et al. 2005), by integrating weather regime information into post-processing methods. Allen et al. (2019) show that the integration of weather regime information into post-processing techniques leads to skill improvements within a highly idealized environment based on the Lorenz 96 model (Lorenz 1996). This work is extended to the operational Global Ensemble Forecasting System from the National Centers for Environmental Prediction in Allen et al. (2020) where weather regime information within post-processing techniques can help to improve skill of wind speed forecasts. The case study from Allen et al. (2020) raises the question of how weather regimes should be defined to maximize its potential to improve forecast skill. This refers to the definition of the weather pattern, including its spatial domain, the methodological description, and whether the weather pattern at model initialization or forecast time should be used within the post-processing methods. Furthermore, Allen et al. (2020) only consider one single extended winter period, while the recent development of year-round weather regimes, as in Grams et al. (2017), allows year-round post-processing techniques, including summer periods.

This article contributes to the scientific discussion described on how to use weather regimes within post-processing techniques to improve NWP forecast skill. We develop a method based on LASSO regression using weather patterns derived from empirical orthogonal function analysis to improve short-term near-surface temperature forecasts in Europe. Within a computational study, the method is compared to classical Moving Average bias corrections. Furthermore, we compare different definitions of weather pattern anomalies with respect to spatial domain (Euro-Atlantic domain vs. region of interest) and weather pattern reference time (weather patterns at NWP initialization vs. NWP realization time). The organization of the article is as follows: We begin with a methodological section (Section 2) that describes how the weather patterns are derived and how the proposed LASSO method can integrate this information. To illustrate how and why this method works, Section 3 explains the rationale behind the use of weather patterns and illustratively describes the method. We then describe the design of the experimental study (Section 4), analyze the results (Section 5), and conclude and discuss them (Section 6).

2. Methodology

a. Weather pattern definition

1) Anomaly calculation

Geopotential height anomalies at 500 hPa are a commonly used meteorological parameter to define weather regimes (Grams et al. 2017; Cassou et al. 2005; Cassou 2008). Anomalies refer to deviations from climatology and are preferable over the raw data because they describe how the weather differs from the typical weather at given time of the year. Most studies using weather regimes focus on single periods, such as the extended winter period (Van Der Wiel et al. 2019; Cassou 2008; Ferranti et al. 2015, 2018) or the summer period (Cassou et al. 2005). More recently, Grams et al. (2017, 2020) proposed a definition of year-round weather regimes by applying data standardization to normalize the geopotential height anomalies. Normalizing geopotential height anomalies is required to consider smaller geopotential height anomalies with lower variability in summer than in winter (Büeler et al. 2021; Wallace et al. 1993). Our geopotential anomaly calculation is loosely inspired by the proposals from Büeler et al. (2021) for weather regime definitions. Note that this study does not use weather regimes but weather patterns that represent the superposition of separate weather signals derived from EOF analysis. The year-round weather regime definition from Grams et al. (2017) uses reanalysis data from past analyses while the definition is extended to forecast data by Büeler et al. (2021). Therefore, we calculate the normalized geopotential height anomalies ϕ_t^a by

$$\phi_t^a = \frac{\phi_t - \bar{\phi}_t}{\sigma_t(\phi)} \tag{1}$$

with ϕ_t geopotential height, $\bar{\phi}_t$ climatological mean geopotential and the standard deviation σ_t as normalization factor. The same formula is used for the calculation of present and forecasted geopotential anomalies. The geopotential heights ϕ_t at model initialization time are derived from the zero-step deterministic forecasts of the open-access TIGGE dataset (Bougeault et al. 2010) based on the average of the semi-daily model runs (0000/1200 UTC). The forecasted geopotential heights are calculated by averaging the forecasts at noon and midnight of the desired lead time meaning that for day-ahead forecasts the lead times 24, 36 and 48 hours are averaged with the purpose to remove high-frequency noise. The calendar day climatologies $\bar{\phi}_t$ are calculated using a 91-day running mean of five-day low-pass filtered geopotential heights (2008-2020) using a Lanczos filter (Duchon 1979). The normalization factor $\sigma_t(\phi)$ is calculated by the 31-day running standard deviation of the geopotential anomalies (2008-2020) and is averaged over the horizontal grid as also described in Büeler et al. (2021). Note that the climatology and the normalization factor are calculated separately for each lead time to account for model drifts (Büeler et al. 2021). This approach differs from Grams et al. (2017) and Büeler et al. (2021) in that only forecasts are used without using reanalysis data. This has the advantage that no second dataset is needed and no additional geopotential forecast calibration is needed.

2) Empirical orthogonal function analysis

Empirical orthogonal functions (EOF) analysis is a common tool in atmospheric science originally introduced by Lorenz (1956). In other scientific domains, it is more commonly known under the term principal component analysis (Wilks 2011) and aims at representing a high-dimensional dataset through a smaller set of independent variables, which are linear combinations of the original ones. As this results in a smaller set of variables, EOF analysis can be regarded, in more modern terms, as a dimensionality reduction technique. When applied to spatio-temporal data, as in this study to geopotential height anomalies, EOF analysis can provide interpretable spatial patterns and their variation over time (Wilks 2011). In the context of atmospheric science, patterns derived from EOF analysis can represent atmospheric oscillations (Wilks 2011) and are therefore regularly used to obtain weather patterns (Michelangeli et al. 1995; Cassou 2008; Grams et al. 2017).

Mathematically, the EOF analysis aims to represent a mean-normalized (anomaly) data matrix $\mathbf{X}$ by calculating the covariance matrix $\mathbf{R} = \mathbf{X}^\mathbf{T}\mathbf{X}$ and solving the eigenvalue problem.

$$\mathbf{RC} = \mathbf{C\Lambda}. \tag{2}$$

$\mathbf{\Lambda}$ is a diagonal matrix containing the eigenvalues $\boldsymbol{\lambda_i}$ and $\mathbf{C}$ contains in its columns the eigenvectors $\mathbf{c_i}$ corresponding to the respective eigenvalue $\boldsymbol{\lambda_i}$. The eigenvectors (EOFs) are usually the core interest of geophysical studies, as when plotted on maps they represent standing oscillations and therefore can be used to derive known physical phenomena such as the North Atlantic Oscillation. The variation of the eigenvectors in time can be calculated by projecting the original data on the eigenvectors

$$\mathbf{a_i} = \mathbf{Xc_i} \qquad \forall i \in p \tag{3}$$

which we refer to as Principal Component time series (PCs) analogous to Björnsson and Venegas (1997). The original data matrix $\mathbf{X}$ can be obtained again through

$$\mathbf{X} = \sum_{i=1}^{p} \mathbf{a_i} \tag{4}$$

with p being the number of EOFs. The eigenvalues of the associated EOFs indicate the amount of variance explained by the respective EOF. The first EOF explains the most variance while each subsequent EOF explains a lower amount of variance and is orthogonal to the EOFs before (Wilks 2011; Björnsson and Venegas 1997). The rationale of EOF analysis is that not all EOFs are required to provide a sufficient representation of the original data matrix $\mathbf{X}$ as few EOFs are often sufficient to explain the dynamical behavior of the system, while higher-indexed eigenvectors often only describe data noise and thus can be ignored. The number of EOFs can be determined based on heuristics or statistical tests such as the North's Test (North et al. 1982). In this study, we restrict the number of EOFs to 14 which is larger than the 7 EOFs in Grams et al. (2017) and Büeler et al. (2021). These 14 EOFs describe 86% (Euro-Atlantic domain) or 97% (Europe) respectively of the data variance. As the subsequent regression method (Section 2b) provides a feature selection, a larger number of EOFs is selected.

3) Weather patterns versus weather regimes

We refer to weather patterns as the result of the EOF analysis of the normalized geopotential anomalies. For the derivation of weather regimes as in Grams et al. (2017), additional steps are needed to derive a discrete set of persistent weather regimes. This comprises the clustering of the PCs using for instance k-means clustering, the derivation of weather regime indices (Michel and Rivière 2011) and the persistence of the regime over a number of days (Grams et al. 2017). Compared to weather regimes, EOFs have the disadvantage of being less interpretable as they are not forced to align with physical processes as atmospheric flows do not follow orthogonal patterns or are uncorrelated (Dommenget and Latif 2002). Due to the prerequisite that all EOFs are orthogonal, later-ranked EOFs are less likely to represent a physical process (Storch and Zwiers 2002). However, PCs are highly informative as early-ranked EOFs are still often physically interpretable and can be used to derive the prevailing atmospheric flows (Ferranti et al. 2018). Furthermore, EOFs provide a highly efficient representation of the data. An EOF is able to represent both states of a prevailing atmospheric flow by describing its negative and positive realizations. For example, an EOF of the North Atlantic Oscillation (NAO) can represent positive and negative NAO indices, while weather regime definitions would require two different classes (Allen et al. 2020). In the context of regression, EOFs have the advantage of avoiding the common issue of multicollinearity as they are uncorrelated by definition.

b. Model Output Statistics with lasso principal component regression

The proposed method uses PCs derived from EOF analysis formulated as a regression problem which is better known in meteorological research as the Model Output Statistics method (Glahn and Lowry 1972). Therefore, deterministic forecasts are calibrated based on the multiple regression formulation

$$\underset{(t\times 1)}{\boldsymbol{b}^{s}} = \underset{(t\times p')}{\boldsymbol{X}} \times \underset{(p'\times 1)}{\boldsymbol{\beta}^{s}} \quad \forall s \tag{5}$$

with the data matrix

$$\boldsymbol{X} = \begin{pmatrix} 1 & a_{11} & \dots & a_{1p'} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & a_{t1} & \dots & a_{tp'} \end{pmatrix}$$

and β_s model parameters. The model is formulated and trained for each site s separately. In the data matrix **X**, the one entries in the first column represent the intercept of the regression, while the other entries **a** represent the chosen PCs derived from Equation 3 with index p' representing the total number of chosen PCs and the intercept. The rows indicate the different time steps t and thus the sample size of the regression problem. The coefficient matrix β contains a weight for each PC and the intercept. In terms of interpretability, the intercept represents a local bias specific to each site, while the weights for the PCs describe the effect of the respective synoptic-scale weather pattern on the specific site s. Note that the matrix X does not suffer from any multicollinearity, since the column of one entries are constants and thus uncorrelated to time-varying PCs, while the PCs are uncorrelated between each other by definition, as earlier described.

The coefficients of Equation 5 could be derived from solving analytically the ordinary least squares problem, as is usually done in Model Output Statistics, but to filter only the relevant PCs, we formulate the regression problem using the LASSO formulation proposed by Hastie et al. (2009) in which the coefficients β are derived from solving the optimization problem

$$\beta = \arg\min_{\beta} \sum_{i \in t} (y_i - \sum_{j \in p'} X_{ij}\beta_j)^2 + \lambda \sum_{j \in p'} |\beta_j| \qquad \forall s. \tag{6}$$

In this formulation, the first summand describes the deviation between observed temperature y and the estimated temperature $X\beta$. The second summand represents the L_1 lasso penalty that leads to sparse solutions of β and thus selects the most relevant weather patterns within the regression. The value of λ controls the amount of regularization with 0 using all weather patterns and ∞ ignoring all weather patterns. We determine the value of λ as a hyperparameter by iterative fitting along a regularization path (Friedman et al. 2010) with a 3-fold cross-validation as implemented in the python-based open-source `scikit-learn` library (Pedregosa et al. 2011). The model is trained on 9 years (2010-2020) of data and evaluated on one left-out year for validation. This scheme, called leave-one-out cross-validation (Wilks 2011), allows to evaluate the model on an independent year to prevent data leakage. The sample size for each model training consists of 9 years of daily data and thus around 3000 samples while 14 PCs are considered in the regression task for the single weather pattern model and 28 PCs for the hybrid models.

Note that using PCs with small eigenvalues does not necessarily imply low relevance within the principal component regression. Examples in the literature show that principal components with a variance explained less than 1% can be highly relevant for the final performance of the regression (Jolliffe 1982).

3. Rationale behind the usage of weather patterns

a. Why using weather patterns?

The calibration of NWP temperature forecasts is often based on simple approaches, such as moving average methods or Kalman filters (Alerskans and Kaas 2021), or based on using exogenous features such as in the Model Output Statistics. Increasingly, Machine Learning techniques (Rasp and Lerch 2018a) are applied to calibrate NWP forecasts based on a large set of different variables and combine them in complex, non-interpretable models. Despite providing highly sophisticated models in terms of complexity, these methods have in common that they focus on site-specific corrections or can only integrate the information from close nearby points using convolutional neural networks (Veldkamp et al. 2021; Li et al. 2022; Cho et al. 2022; Xiang et al. 2022). The information of weather patterns, similar to the usage of weather regimes in Allen et al. (2019, 2020), provides an elegant way to include flow-dependent information by using a small set of scalars that describe the synoptic-scale flow.

There are different weather patterns that can be informative within a regression context. Figure 1 illustrates the different possibilities of geopotential anomalies considered in this study that differ in terms of spatial domain (large-scale flow versus region of interest) and reference time (model initialization versus forecast reference time). Using large-scale Euro-Atlantic weather patterns is the most common approach as it allows to identify large-scale atmospheric flows that largely influence the weather in the mid-latitudes. These are typically used for the definitions of weather regimes (Grams et al. 2017) and have a large-scale spatial domain (30°N to 90°N, 80°W to 40°E) as in Grams et al. (2017) which is also used for the Euro-Atlantic weather pattern definition in this study. The second spatial domain considered is the region-of-interest domain in which the

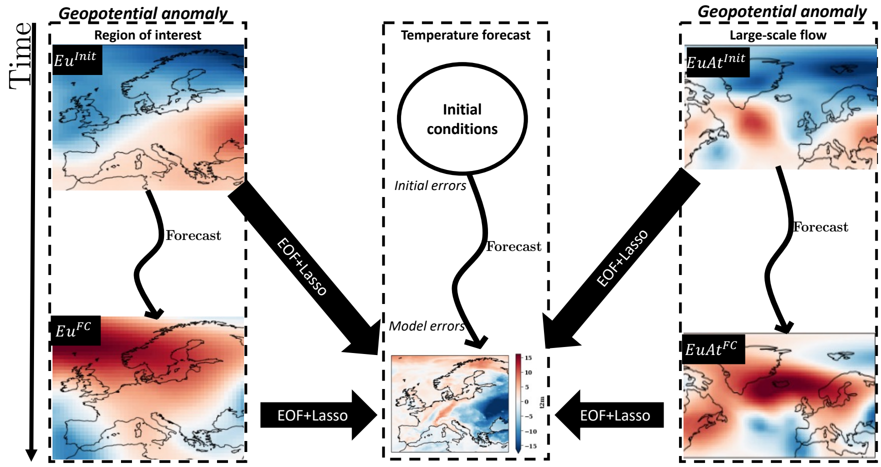

FIG. 1. Illustration of how different weather pattern definitions are used in the proposed method. The example illustrates different geopotential anomalies in terms of spatial domain and reference time for the case of predicting temperature forecast errors with a lead time of 7 days.

spatial domain of the weather pattern is identical to the spatial domain of the temperature forecasts (33°N to 72°N, 12°W to 35°E). This approach assumes that the large-scale atmospheric flow is less relevant than the actual impact of different atmospheric flows on the region. In addition to the spatial domain, we investigate the importance of the reference time of the weather pattern. Weather prediction is an initial value problem (Bjerknes et al. 2009) meaning that the future state of the atmosphere can be predicted by knowing the initial state of the atmosphere and the application of physical laws. NWP models are known to have initial errors and model errors (Rodwell et al. 2018) which is why both can potentially be relevant to estimate a weather pattern bias. Using weather patterns at model initialization time ($\mathrm{Eu}^{\mathrm{Init}}$ and $\mathrm{EuAt}^{\mathrm{Init}}$) can consider systematic forecast errors at model initialization time, while using weather patterns at forecast time ($\mathrm{Eu}^{\mathrm{FC}}$ and $\mathrm{EuAt}^{\mathrm{FC}}$) can consider systematic model errors at different model lead times. The importance of these different settings is evaluated in the computational section of this study.

b. Illustration of the proposed method

To better understand the proposed method, Figure 2 shows for three consecutive days the geopotential height anomaly, the output of a bias-corrected temperature forecast (45-day Moving Average), and the estimated bias correction of the proposed method. The bias correction aims to

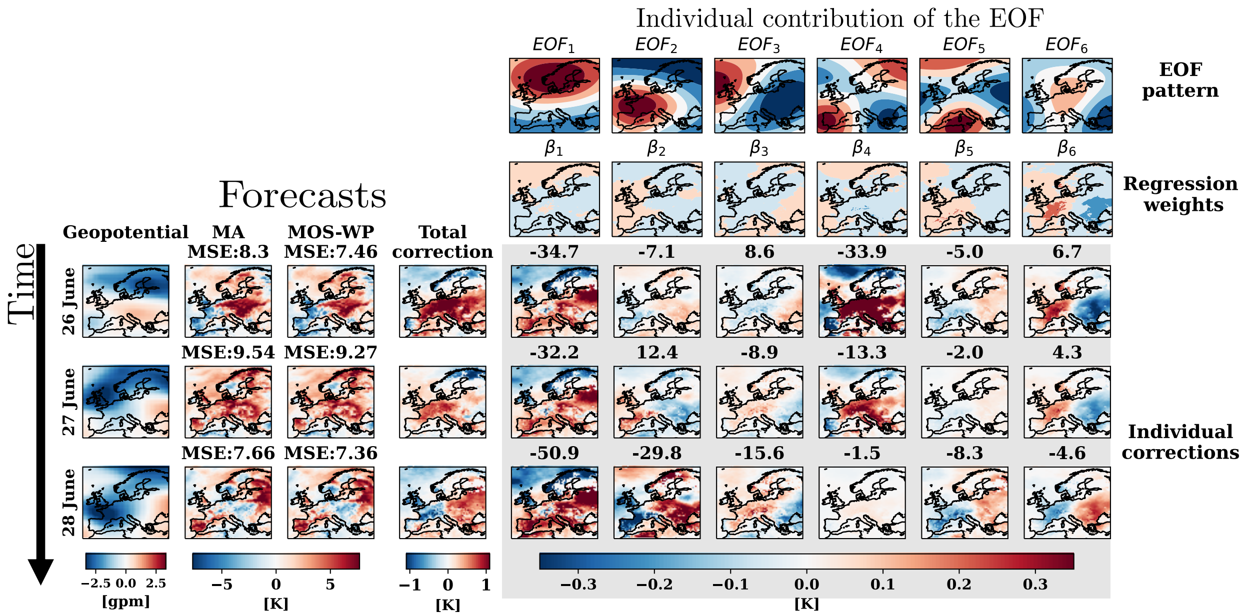

FIG. 2. Illustration of the proposed method. The temporal period (rows 3-5) spans three days (26 June 2017 to 28 June 2017). The gray box illustrates the changes induced by the respective EOF (represented in the columns). The titles of the contour plots in the gray box indicate the PCs for a given time (row) and the EOF pattern (column). MA represents the temperature forecast error after applying the Moving Average method, MOS-WP the forecasts after bias correction.

systematically estimate the forecast error after the application of the Moving Average method to isolate the induced forecast errors of the weather pattern. The gray box exhibits the PCs, while the columns refer to the different EOFs. For each EOF, the EOF pattern (correlation with the geopotential anomaly), the calculated regression weights, and the individual contributions are shown. The total correction can be calculated by the sum of the products between PC and the regression weights.

On 26 June 2017, there is a large positive geopotential anomaly in Eastern Europe and a smaller cyclonic pattern over Western Europe. The EOFs decompose this geopotential anomaly into different weather patterns, while the first and fourth EOFs are the most relevant to reconstruct the geopotential anomaly, as evident in the largest PC values. The first EOF detects the separation in Northern and Southern Europe with a large negative PC value (-34.7) indicating a negative realization of the first EOF. As the geopotential anomaly is characterized by an additional low-pressure anomaly over Iberian Peninsula, a second EOF is needed besides the distinction of pressure anomalies in Southern and Northern Europe (EOF1). The negative realization of EOF4 addresses this by detecting four distinct pressure centers, each with the same pressure sign, arranged diagonally. This weather situation is an example of a weather situation that cannot be assigned to a single EOF showing that this is not a highly recurring and stable pattern. In terms of the bias correction, the temperature forecasts are overestimated in Central Europe which align with the forecasted bias during this situation. The individual contribution at this time step can be derived by multiplying the weights by the PCs. The total correction of the method is then derived by summing all individual contributions in alignment with Equation 5. In this weather situation, the main contribution comes from the fourth EOF indicated by the strong values in the map and the similarity of its individual correction and the total correction. By comparing the second and third columns, the reduction of the temperature overestimation is noticeable, as indicated by a smaller mean square error (7.46 versus 8.3). During the next two days, the weather system changes to a more pronounced low-pressure system over the United Kingdom that strongly resembles the negative loadings of EOF2. This leads to a negative temperature forecast bias in Western Europe and an overestimation in Eastern Europe that the proposed bias correction method is able to correct as visible in the lower MSE.

In addition to illustrating how the method works, this example also shows how quickly the bias correction can adapt to changing weather conditions, leading to different bias corrections in a short time.

4. Experimental design

a. Data

The short-term temperature forecasts originate from the deterministic high-resolution ECMWF Integrated Forecasting System (IFS) initialized daily (1200 UTC) with a lead time of day-ahead (+24h) to week-ahead (+168h). The forecasts are remapped to a grid spacing of 0.25° ($\approx$ 31 km) to match the ERA5 reanalysis model (Hersbach et al. 2020) used for forecast verification. The geopotential heights are derived from the open-access TIGGE dataset (Bougeault et al. 2010) and the respective anomalies are calculated as described in Section 2.

b. Computational experiments

The aim of the computational experiments is to determine the value of the different weather patterns, to evaluate the improvements in forecast skill, and to determine whether the proposed method improves the meteorological understanding that leads to these biases. For this purpose, we compare six different weather pattern definitions with differences in their spatial domain and temporal definitions. The spatial domains investigated are the European (Eu, 33°N to 72°N, 12°W to 35°E) and the Euro-Atlantic (EuAt, 30°N to 90°N, 80°W to 40°E) domains, while the forecast reference time refers to using weather patterns at model initialization (Init) or forecast (FC) time. This leads to four different weather pattern definitions ($\mathrm{Eu}^{\mathrm{Init}}$, $\mathrm{Eu}^{\mathrm{FC}}$, $\mathrm{EuAt}^{\mathrm{Init}}$ and $\mathrm{EuAt}^{\mathrm{FC}}$) as also described in Section 3. Furthermore, two additional hybrid definitions are proposed consisting of the combination of both spatial domains with forecasted weather patterns ($\mathrm{Eu}^{\mathrm{FC}}$ & $\mathrm{EuAt}^{\mathrm{FC}}$) and one definition based on the combination of the two temporal reference times ($\mathrm{Eu}^{\mathrm{Init}}$ & $\mathrm{Eu}^{\mathrm{FC}}$) for the European domain. Using both spatial domains is motivated by the multi-scale behavior of the atmosphere while using weather patters at model initialization and forecast time allows to systematically correct biases associated with both model initialization and forecast time. This leads in total to six different weather pattern definitions. The methods with individual weather patterns contain the 14 leading EOFs, whereas the hybrid methods contain 14 EOFs from the respective individual weather pattern, and thus consider 28 EOFs. Despite this gives the hybrid methods an advantage due to more information, it allows to measure the value of the hybrid methods.

The performance of the methods is benchmarked against a Moving Average bias correction with a window length of 45 days that is calculated for each lead time separately. Although there has been a wide range of more advanced methods, for instance, methods applying Machine Learning with a large number of additional variables (Rasp and Lerch 2018a), Moving Averages are still commonly used for bias correction as they are simple to implement and perform well against other methods such as Kalman filters (Alerskans and Kaas 2021). A window length of 45 days is selected, as moving training window lengths between 30 and 45 days have shown good skill in the literature for temperature forecast postprocessing (Gneiting et al. 2005). To directly evaluate whether the proposed method improves compared to the Moving Average method, we select the Mean Squared Error Skill Score (MSESS) which is a metric to evaluate deterministic forecasts in comparison to deterministic reference forecasts. For more information on the MSESS, the reader is referred to the appendix.

To validate the method, the dataset is split in a cross-validation fashion into a training and a test data set, while the training period covers the years 2010-2020 excluding one left-out test year to evaluate the method. This procedure is repeated until each year between 2013 and 2018 was used as a test year. As a separate model is fitted for each year, it also provides information about the robustness of the model.

5. Results

a. Weather pattern definitions

First, we compare the different definitions of the weather patterns to identify the most suitable weather pattern definition. Figure 3 shows the spatially averaged MSESS for the six different weather pattern definitions. All investigated weather patterns show skill improvements at shorter lead times with highest MSESS at day-ahead forecasts and a subsequent decline of MSESS to a lead time up to four days. The best performing method is the $\mathrm{Eu}^{\mathrm{FC}}$ method that uses weather patterns with the same spatial domain as the temperature forecasts and also the same valid time. This method strictly outperforms all other methods with single weather pattern definitions for all lead times showing that it is the preferred weather pattern to be included in the proposed method. In numbers, $\mathrm{Eu}^{\mathrm{FC}}$ provides skill improvements between 1.5 and 3%.

Using forecasted weather patterns shows superior skill than weather patterns at initialization time. This is evident when comparing the performance of the methods using the same spatial domain but different reference times (e.g. $\mathrm{Eu}^{\mathrm{FC}}$ and $\mathrm{Eu}^{\mathrm{Init}}$). The informative value of the weather pattern at model initialization time shows no skill improvements after four days lead time while the $\mathrm{Eu}^{\mathrm{FC}}$ even shows an inflection point after five days lead time. With respect to the spatial domain, weather patterns with the same spatial domain as the forecasts are preferable over weather patterns that describe the large-scale atmospheric flow. This is evident in the constantly better skill scores for $\mathrm{Eu}^{\mathrm{FC}}$ with around 1% better skill scores than for $\mathrm{EuAt}^{\mathrm{FC}}$.

The only method that can compete for up to two days lead time with $\mathrm{Eu}^{\mathrm{FC}}$ is the hybrid method $\mathrm{Eu}^{\mathrm{Init}}\&\mathrm{Eu}^{\mathrm{FC}}$ that also contains the forecasted region-of-interest weather pattern. Additionally, the other hybrid method $\mathrm{Eu}^{\mathrm{FC}}\&\mathrm{EuAt}^{\mathrm{FC}}$ also contains the $\mathrm{Eu}^{\mathrm{FC}}$ weather pattern and shows inferior

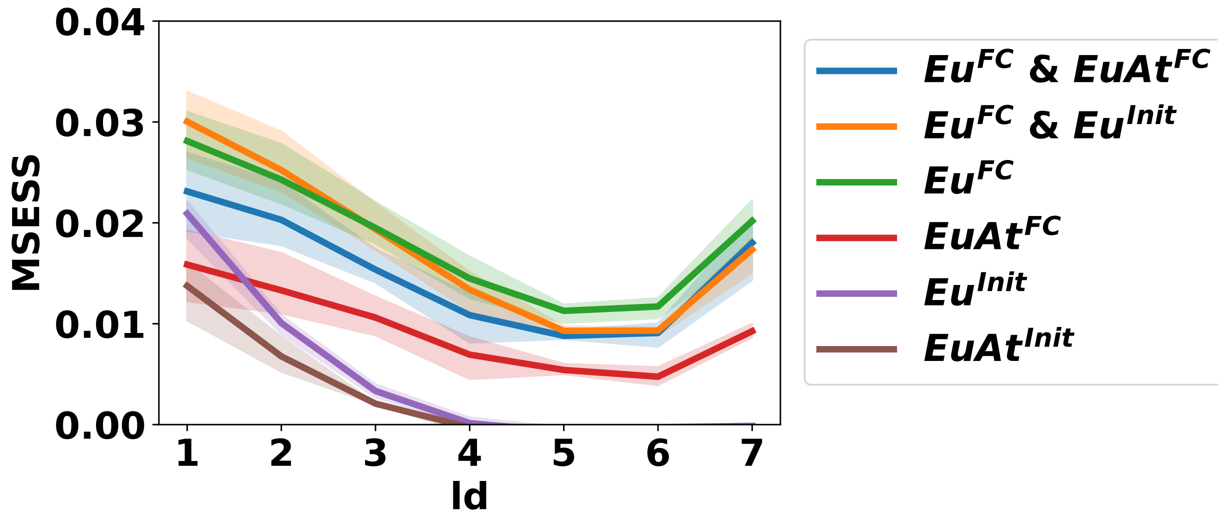

FIG. 3. Forecast skill of different methods validated based on the MSESS over all sites including sea surfaces. The confidence intervals are created by showing the worst, median and the best performance of the years 2013, 2015 and 2017. The first term of the model names indicate the spatial domain of the weather pattern, while the superscript refers to the weather pattern reference time.

performance. This indicates that it becomes more challenging for the LASSO algorithm to identify the most relevant PCs with an increasing number of PCs. Therefore, we select the $\mathrm{Eu}^{\mathrm{FC}}$ for the subsequent analyses.

b. Performance

Figure 4 shows the yearly and monthly performance spatially averaged over the spatial domain of the temperature forecasts. All yearly and monthly MSESS values are positive, indicating that the proposed method is able to provide year-round and monthly performance improvements. Between the investigated years, the performance is relatively similar, showing skill improvements of around 2% to 3% in the day-ahead range. Day-ahead forecasts show the greatest improvement of all lead times. With longer lead times, skill improvements deteriorate and then increase again (> 6 days). The monthly distribution of skill improvements shows that the spring and autumn months have highest skill improvements. More specifically, during spring, skill scores of more than 5% are observable between one and three days lead time. Lower skill scores are particularly noticeable during the winter months.

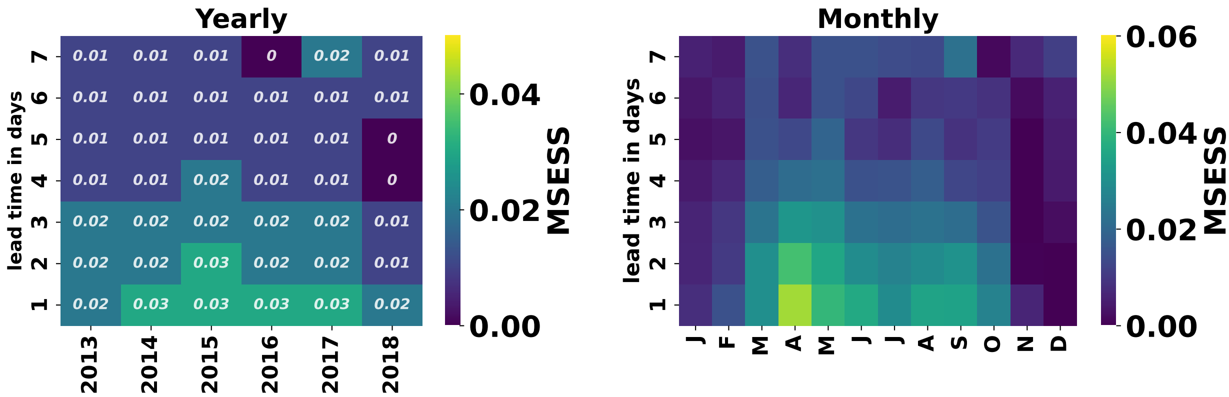

FIG. 4. Forecast performance for different years and for different months. The monthly averages are calculated over all test years.

The spatial distribution of skill improvements (Figure 5) shows which regions benefit most using the proposed bias correction. Similarly to monthly and yearly averages, the method shows consistent improvements. Skill improvements on the land surface are much greater than on the sea surface. This can be explained by the higher thermal inertia of the sea surface that reduces the local response rate on weather patterns. At short lead times, a large number of sites greatly benefit from the proposed method. On land surfaces, skill improvements exceeding 6% are observable. Furthermore, there are differences in local skill scores with respect to the lead time. At shorter lead times (d+1 to d+3), most of Europe (excluding Scandinavia) show high skill scores, while at longer lead times highest skill scores are observable in Southern Europe. With longer lead times, forecast improvements become larger and smoother leading to large-scale forecast skill improvements in Eastern Europe at week-ahead temperature forecasts.

c. Method interpretability

The proposed method has the advantage that it is highly interpretable due to the interpretability of the EOF analysis and the determined regression weights. Therefore, we analyze whether this simplicity can help us derive insights into the meteorological significance of the results. Figure 6 depicts the mean contribution of the respective EOF which explains how much each EOF contributes to the bias correction averaged over all time steps. The explained cumulative variance shows how much of the total data variance is explained by the respective EOFs while the autocorrelation provides information about the persistence of the weather patterns. Using the first

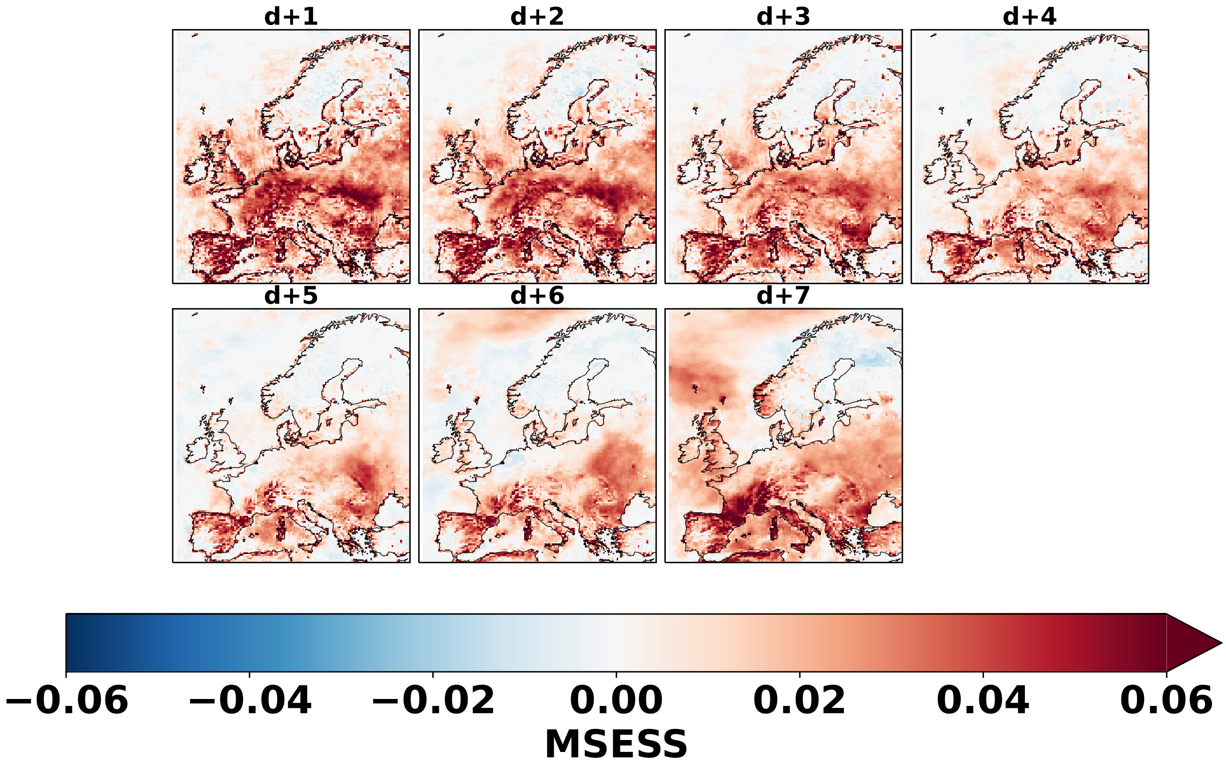

Fig. 5. Spatial distribution of the MSESS for different lead times averaged over all test years. The value range has been reduced to only showing values up to 0.06 to make the results clearer.

seven PCs, about 90% of the data variance of the geopotential anomalies can be explained, which highlights the importance of these first seven weather patterns. The first three weather patterns explain around 68% of the data and contribute the most to the bias correction at short lead times as noticeable by their large relative mean contribution. At longer lead times, the bias correction relies on a larger number of PCs illustrated by the more uniform contributions of the different PC mean contributions. Interestingly, PCs which do not explain much variance can become important for the regression visible by the relatively large importance of PCs indexed between 7 and 10 for week-ahead bias correction. This may seem counterintuitive, but aligns with the observation in Jolliffe (1982) that low explained variance of the EOF analysis does not imply low importance in a regression context.

Early-indexed EOFs are more persistent as visible in larger autocorrelations, whereas later-indexed EOFs are a lot less steady and thus indicate a more variable weather situation. To better understand what these EOFs represent, Figure 7a depicts the first seven EOFs and Figure 7b

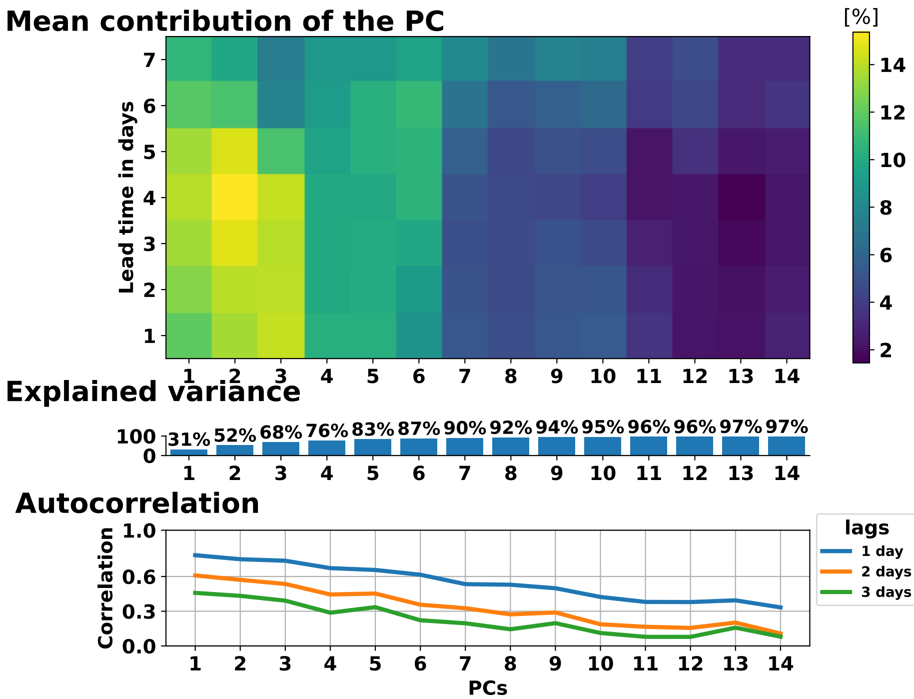

FIG. 6. Derived mean absolute change of temperature associated with the respective PC, the associated cumulative explained variance and the autocorrelation for selected lag days. The mean contribution is normalized based on total contribution given lead time (rows).

the regression weights β associated with the respective EOFs at different lead times. The EOFs represent the correlation between the PCs and the geopotential anomaly, meaning that positive (negative) values of the EOF pattern align with positive (negative) geopotential anomalies.

The EOF patterns can be aligned with the derived weather regimes in Grams et al. (2017). The first derived EOF (31% explained variance) is characterized by a strong positive anomaly in Scandinavia and minor negative anomalies in Southern Europe. This pattern resembles blocking patterns, such as the known variants of the Scandinavian Blocking and European Blocking (Grams et al. 2017). As the centers of these blocking patterns are close to each other, it is likely that the first EOF can represent the meteorological signal of both weather regimes while ignoring the assumption of weather regime persistence. In its negative realization, this EOF has a large-scale low-pressure anomaly over Scandinavia that is similar to the Scandinavian Trough. The second EOF is characterized by a strong pressure anomaly located farther south than in the first EOF located in Western Europe. In the context of Euro-Atlantic weather regimes, the EOF resembles

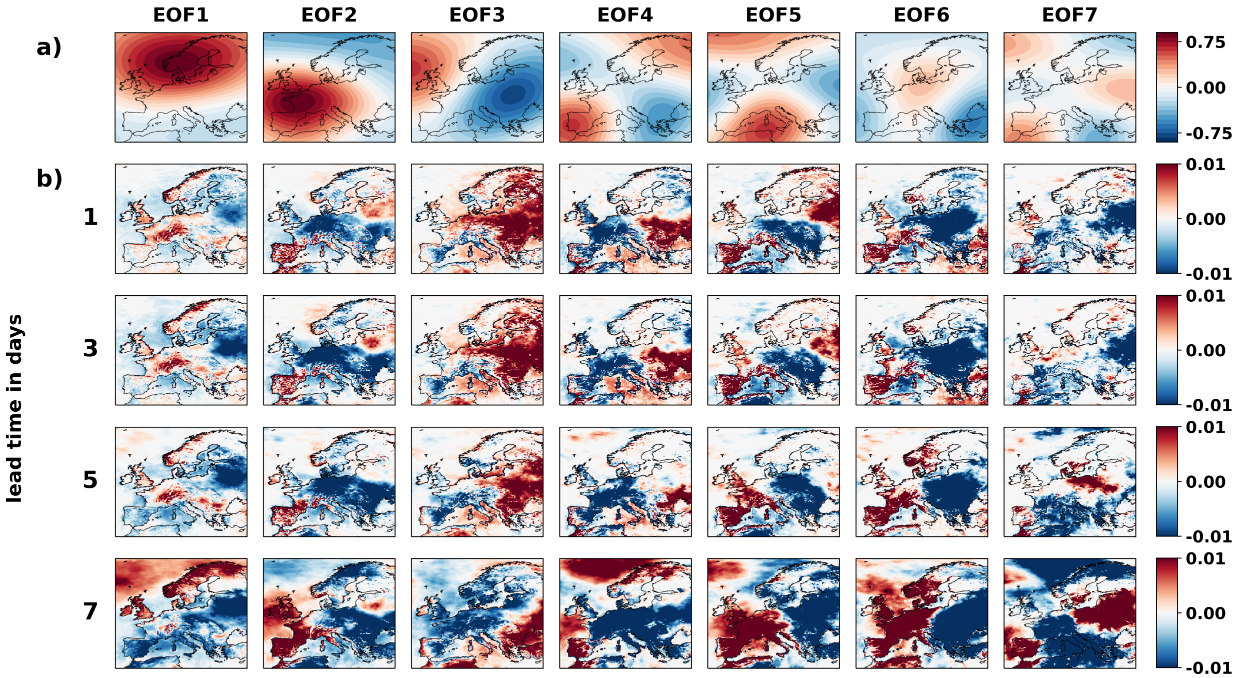

FIG. 7. Relationship between the spatial distribution of the EOF (a) and the regression weights determined by the LASSO regression. The spatial pattern of the EOF is calculated based on the correlation between the PC and the geopotential anomaly. The rows contain the same color scale. Note that the EOFs (a) are selected based on the lead time of 7 days as there are no visible differences between the EOF output for different lead times. In case of an inverse relationship, the sign of the weights is modified to match the EOF pattern.

weather regimes characterized by a ridge in Southern Europe such as during the Zonal regime. In its negative realization, the EOF2 pattern resembles the Greenland blocking weather regime. The third EOF shows a dipole pattern with locations in the Atlantic Ocean and in Eastern Europe. This EOF is similar to the weather regimes Atlantic Ridge and Atlantic trough as derived in Grams et al. (2017). Given the high persistence, the large-scale structure of the pressure system, and the similarity to known weather regimes, we argue that the first three weather patterns describe to a significant amount the information of weather regimes without making assumptions about the persistence and signal strength. Later-indexed EOFs are more likely to represent unstable weather patterns such as during the transition between more stable pressure systems.

The regression weights (Figure 7b) can be used to identify the regions most affected by the respective EOF. As the bias correction is the product of the static regression weights and the strength of the respective EOF, the stronger the EOF, the larger the bias correction of the proposed method (Figure 7b) with the same sign as the respective EOF. A negative (positive) bias sign indicates a systematic NWP temperature underestimation (overestimation) and, as the bias is subtracted, a bias correction toward higher (lower) temperatures. For the interpretation of Figure 7b), negative (positive) values lead to larger (smaller) temperatures in the sign of the EOF, while the negative realization of the EOF leads to the opposite.

An interesting meteorological situation for forecasting are blocking events as they are known to be prone to forecast errors (Ferranti et al. 2015). Blocking events are well-researched, long-lasting weather regimes that prevail the westerly atmospheric flow in Europe and are associated with higher temperatures in summer and lower temperatures in winter (Grams et al. 2017). In terms of bias correction, when large-scale blocking patterns occur over Scandinavia (EOF1), the proposed method systematically corrects overestimated temperatures in selected regions of the United Kingdom, Norway, and the Alpine region. In case of a large high-pressure anomaly in Western Europe (EOF2), the model systematically corrects underestimated temperatures in Central Europe. Interestingly, this aligns with the analysis from Lemburg and Fink (2022) that show negative biases during blocking events for daily 2m maximum temperature forecast errors at a lead time of 3 days for the ECMWF-IFS ensemble. The derived biases in Lemburg and Fink (2022) strongly resemble the derived weights in this study at three days ahead lead time. At longer lead times, the underestimation moves towards Eastern Europe, while in the center of the blocking pattern in Western Europe temperature overestimations are corrected.

When comparing the regression weights at week-ahead lead time with the EOF pattern, there is a clear similarity in sign and spatial distribution. Higher geopotential heights do not necessarily imply higher temperatures at the same site due to changes in atmospheric flows, yet often have a similar effect close to the geopotentials. This aligns with the temperature anomalies for different weather regimes as analyzed in Grams et al. (2017). When comparing the regression weights of the first three EOFS with the respective temperature anomalies of the associated weather regimes in Grams et al. (2017), there is a high similarity between both patterns. This validates the assumption that a bias-specific temperature anomaly is learned by the method. As the sign of the bias correction overlaps with the geopotential anomaly, it also means that at longer lead times the proposed model systematically corrects the NWP forecast by typical observable temperatures for the forecasted weather pattern. Therefore, the proposed bias correction model makes the NWP forecast less keen at longer lead times. At shorter lead times, the proposed weights are more diverse and often show the opposite sign of the EOF pattern. This means that in these situations, the proposed model uses the forecasted EOF pattern to make the forecasts bolder. For example, at short lead times the bias correction systematically corrects underestimated impacts in the Alpine region during blocking events over Scandinavia (EOF1), increases the temperatures during blocking events in Central Europe (EOF2) or decreases temperature during large cyclonic patterns in Eastern Europe (EOF3).

A plausible explanation why the model corrects at longer lead times typical temperature anomalies of the weather pattern can be explained by the reliability of the geopotential forecast. The proposed method only uses a single geopotential forecast meaning that no uncertainty of the geopotential forecast is considered. Single geopotential forecasts quickly become less reliable, as noticeable in large weather regime ensemble spreads and strong declines of the probabilistic predictability of weather regimes after a few days (Büeler et al. 2021). Therefore, a single forecast for the weather pattern is not reliable at longer lead times making the learning of associated biases also less reliable. At short lead times, the geopotential forecasts are highly reliable, which means that the bias correction model uses the property that large-scale atmospheric flows have higher predictability than high-frequency variability (Lorenz 1969). This allows to learn specific biases associated with the weather pattern. Therefore, a plausible explanation for the highlighted inflection points is the point at which the the proposed method does not trust a single realization of the weather pattern, and instead attempts to mitigate the specific impact of the forecasted weather pattern on the temperature. As the forecasts in this study originate from a single high-resolution NWP model, this effect is similar to the common observation that ensemble means have higher skill than single forecasts in particular at longer lead times.

6. Discussion and conclusion

This study is motivated by recent studies showing the descriptive ability of weather regimes to explain meteorological situations associated with higher forecast errors (Ferranti et al. 2015), year-round weather regime definitions (Grams et al. 2017; Büeler et al. 2021) and recent advances in the integration of weather regime information into forecast calibration methods (Allen et al.

2019, 2020). We propose a bias correction method inspired by Model Output Statistics which embeds the information of weather patterns inside a LASSO regression to improve NWP forecast skill of deterministic short-term temperature forecasts. Instead of using weather regimes, we use weather patterns as the output from the EOF analysis that are easier to derive, contain more information than weather regimes and can express highly variable weather pattern situations. Instead of using weather regimes, we use weather patterns as the output from the EOF analysis that are easier to derive, contain more information than weather regimes and can express highly variable weather pattern situations. Moreover, preliminary tests (not shown) suggested that regime-based post-processing was not able to reach the same level of predictive skill as the pattern-based technique described here. We show, for the first time, that temperature forecasts can greatly benefit from this information, since spatially averaged skill improvements up to 3% are observable with significantly larger values over the land surface. We show that the proposed methodology shows the best performance during spring and autumn months, but the method achieves year-round skill improvements. This emphasizes the importance of investigating entire years in forecast calibration studies in contrast to focusing on single periods such as the well-researched (extended) winter period (Ferranti et al. 2015; Allen et al. 2019, 2020; Barnes et al. 2019).

Subject to the lead time, the performances of the methods show a U-shaped pattern with high skill scores at short lead times, decreasing scores afterwards until inflection points of skill scores are reached at around lead times of five or six days subject to the respective site. Based on an analysis of the EOF patterns and its respective regression weights, we show systematic differences of the model weights between short and long lead times. At short lead times, the model is confident in making forecasts as illustrated by the model capability to correct systematic negative biases during blocking events in Europe. At longer lead times, the model systematically reduces the weather pattern induced temperature anomaly making the forecast less bold. We argue that this originates from the lower reliability of deterministic geopotential forecasts. Therefore, a promising further research avenue is to incorporate the weather pattern uncertainty by using multiple geopotential height forecasts from an ensemble forecasting system, for instance based on the works from Büeler et al. (2021) to derive weather regime probabilities. As the spread of the ensemble of forecasts is related to the prevailing atmospheric flow (Rodwell et al. 2018), there is also further research needed to evaluate the value of weather patterns for temperature ensemble forecasts. This is particularly true for studies with a focus on longer lead times than those investigated in this study as the uncertainty increases and deterministic geopotential and temperature forecasts are not reliable.

Furthermore, we compared six different weather pattern definitions in terms of their spatial domain and the weather pattern reference time. Higher skill scores are observable in case the EOF analysis uses weather patterns with a spatial domain identical to the forecast domain. This supports the hypothesis formulated in Allen et al. (2020) that specialized regime definitions can provide more value within post-processing techniques than large-scale regime definitions. We approve this statement by showing that weather pattern definitions with common large-scale spatial domains, as in Grams et al. (2017) and Ferranti et al. (2015), show less skill than the weather patterns with the same spatial domain as the forecast despite these contain more information about the large-scale atmospheric flow. Further research avenues could investigate whether this statement is also valid for other meteorological variables such as wind speed or precipitation. Furthermore, we show that using forecasted weather patterns is more skillful than using weather patterns at the model initialization time. Using hybrid models did not lead to models that outperform the single $\mathrm{Eu}^{\mathrm{FC}}$ model highlighting the importance of forecasted weather patterns and the region-of-interest spatial domains. Note that in the regression of this study, the hybrid models only contain individual terms without interaction terms between the weather patterns. Therefore, an interesting research direction is the inclusion of interaction terms between initial and forecasted weather patterns, which allows one to obtain individual weights for weather pattern trajectories. Note that this dramatically increases the number of predictors, which is in the case of 14 EOFs an additional number of 196 interaction terms (e.g. for $\mathrm{Eu}^{\mathrm{Init}}$ & $\mathrm{Eu}^{\mathrm{FC}}$: $n_{EOF^{Init}} \times n_{EOF^{FC}}$).

Finally, the EOF analysis used in this study offers a straightforward method to incorporate information on the large-scale atmospheric flow into post-processing techniques. Most post-processing techniques are applied point-wise, and thus do not allow the integration of information about the atmospheric flow situation. This is also the case for recent proposals based on more sophisticated Machine Learning methods, such as those described in Rasp and Lerch (2018b). Methodologically, it is possible to include spatial information in neural networks through the usage of Convolutional Neural Networks. In these architectures, convolutional operations based on learnable kernels enable the aggregation of information from neighboring sites. This has shown promise for NWP post-processing techniques for wind speeds (Veldkamp et al. 2021), precipitation

(Li et al. 2022) and also air temperatures (Cho et al. 2022; Xiang et al. 2022). Although in principle this enables the integration of spatial information into neural networks, kernels are limited in their receptive field by their kernel size. More concretely, the effective receptive field is proportional to $O(K\sqrt{L})$ with K kernel size and L stacked layers (Luo et al. 2016). Typical kernel sizes are small, with typical sizes of 3 by 3 pixels. This raises questions about the ability of CNNs to effectively capture and integrate synoptic-scale information from atmospheric flows. Consequently, it would be worthwhile to explore whether neural networks could gain additional benefits from incorporating atmospheric flow information through forecasted weather patterns as derived in this study.

Acknowledgments. The presented work has been carried out within the national research project "SOLREV" (FKZ 03EE1010E) funded by the Federal Ministry for Economic Affairs and Climate Action (BMWK) on the basis of a decision by the German Bundestag. M. Zech has been supported by the German Federal Environmental Foundation within the PhD scholarship (grant no. 20020/667-33/2).

Data availability statement. The data required to reproduce the weather patterns is publicly available as the open access TIGGE dataset (`https://apps.ecmwf.int/datasets/data/tigge/levtype=sfc/type=cf/`). The reanalysis dataset ERA5 can be accessed via the Climate Data Store `https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=form`. The temperature forecast can be downloaded from the ECMWF IFS after registration. All scripts to run the experiments are provided in a publicly available git repository.

APPENDIX

The forecasts are evaluated based on skill scores which relate the evaluation metrics to a reference forecast as commonly applied in forecasting applications. The reference forecast applied in this study is a bias correction which is applied in a rolling training window fashion. The bias corrected forecasts term y_{bc}, the forecasts based on the proposed method y_{pred} and the actual temperature measurements y.

The Mean Absolute Skill Score (MAESS) formulates

$$MAESS = 1 - \frac{MAE}{MAE^{ref}} = 1 - \frac{T^{-1}\sum_t |y_t^{pred} - y_t|}{T^{-1}\sum_t |y_t^{bc} - y_t|} \tag{A1}$$

which measure the absolute distance. This metric assumes that all deviations from the analysis should get the same weight which as a second metric this study uses the, more widely used, Mean Squared Skill Score (MSESS)

$$MSESS = 1 - \frac{MSE}{MSE^{ref}} = 1 - \frac{T^{-1}\sum_t (y_t^{pred} - y_t)^2}{T^{-1}\sum_t (y_t^{bc} - y_t)^2} \tag{A2}$$

which penalized larger errors disproportionately more. In case of temperature, this metric makes sense as larger errors are usually more crucial than smaller ones.

References

Alerskans, E., and E. Kaas, 2021: Local temperature forecasts based on statistical post-processing of numerical weather prediction data. *Meteorological Applications*, **28 (4)**, https://doi.org/10.1002/met.2006.

Allen, S., C. A. Ferro, and F. Kwasniok, 2019: Regime-dependent statistical post-processing of ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, **145 (725)**, 3535–3552, https://doi.org/10.1002/qj.3638.

Allen, S., C. A. Ferro, and F. Kwasniok, 2020: Recalibrating wind-speed forecasts using regime-dependent ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society*, **146 (731)**, 2576–2596, https://doi.org/10.1002/qj.3806.

Barnes, C., C. M. Brierley, and R. E. Chandler, 2019: New approaches to postprocessing of multi-model ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, **145 (725)**, 3479–3498, https://doi.org/10.1002/qj.3632.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525 (7567)**, 47–55, https://doi.org/10.1038/nature14956.

Bjerknes, V., E. Volken, and S. Brönnimann, 2009: The problem of weather prediction, considered from the viewpoints of mechanics and physics. *Meteorologische Zeitschrift*, **18 (6)**, 663–667, https://doi.org/10.1127/0941-2948/2009/416.

Björnsson, H., and S. A. Venegas, 1997: A manual for EOF and SVD analyses of climatic data. Tech. Rep. 1, CCGCR, 112–134 pp. https://doi.org/10.1080/01422419908228843, URL http://www.geog.mcgill.ca/gec3/wp-content/uploads/2009/03/Report-no.-1997-1.pdf%5Cnhttp://doi.wiley.com/10.1029/2002JD002670, 9809069v1.

Bougeault, P., and Coauthors, 2010: The thorpex interactive grand global ensemble. *Bulletin of the American Meteorological Society*, **91 (8)**, 1059–1072, https://doi.org/10.1175/2010BAMS2853.1.

Büeler, D., L. Ferranti, L. Magnusson, J. F. Quinting, and C. M. Grams, 2021: Year-round sub-seasonal forecast skill for Atlantic–European weather regimes. *Quarterly Journal of the Royal Meteorological Society*, **147 (741)**, 4283–4309, https://doi.org/10.1002/qj.4178.

Cassou, C., 2008: Intraseasonal interaction between the Madden-Julian Oscillation and the North Atlantic Oscillation. *Nature*, **455 (7212)**, 523–527, https://doi.org/10.1038/nature07286.

Cassou, C., L. Terray, and A. S. Phillips, 2005: Tropical Atlantic influence on European heat waves. *Journal of Climate*, **18 (15)**, 2805–2811, https://doi.org/10.1175/JCLI3506.1.

Cho, D., C. Yoo, B. Son, J. Im, D. Yoon, and D. H. Cha, 2022: A novel ensemble learning for post-processing of NWP Model's next-day maximum air temperature forecast in summer using deep learning and statistical approaches. *Weather and Climate Extremes*, **35**, 100 410, https://doi.org/10.1016/j.wace.2022.100410, URL https://doi.org/10.1016/j.wace.2022.100410.

Dommenget, D., and M. Latif, 2002: A cautionary note on the interpretation of EOFs. *Journal of Climate*, **15 (2)**, 216–225, https://doi.org/10.1175/1520-0442(2002)015⟨0216:ACNOTI⟩2.0.CO;2.

Duchon, C. E., 1979: Lanczos Filtering in One and Two Dimensions. *Journal of applied meteorology*, **18 (8)**, 1016–1022, https://doi.org/10.1175/1520-0450(1979)018⟨1016:LFIOAT⟩2.0.CO;2.

Ferranti, L., S. Corti, and M. Janousek, 2015: Flow-dependent verification of the ECMWF ensemble over the Euro-Atlantic sector. *Quarterly Journal of the Royal Meteorological Society*, **141 (688)**, 916–924, https://doi.org/10.1002/qj.2411.

Ferranti, L., L. Magnusson, F. Vitart, and D. S. Richardson, 2018: How far in advance can we predict changes in large-scale flow leading to severe cold conditions over Europe? *Quarterly Journal of the Royal Meteorological Society*, **144 (715)**, 1788–1802, https://doi.org/10.1002/qj.3341.

Friedman, J., T. Hastie, and R. Tibshirani, 2010: Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33 (1)**, 1–22, https://doi.org/10.18637/jss.v033.i01.

Glahn, H. R., and D. A. Lowry, 1972: The Use of Model Output Statistics (MOS) in Objective Weather Forecasting. *Journal of Applied Meteorology*, **11 (8)**, 1203–1211, https://doi.org/10.1175/1520-0450(1972)011⟨1203:tuomos⟩2.0.co;2.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, **133 (5)**, 1098–1118, https://doi.org/10.1175/MWR2904.1.

Grams, C., L. Ferranti, and L. Magnusson, 2020: How to make use of weather regimes in extended-range predictions for Europe. *ECMWF Newsletter*, **No. 165 (165)**, 14–19, URL https://www.ecmwf.int/en/newsletter/165/meteorology/how-make-use-weather-regimes-extended-range-predictions-europe.

Grams, C. M., R. Beerli, S. Pfenninger, I. Staffell, and H. Wernli, 2017: Balancing Europe's wind-power output through spatial deployment informed by weather regimes. *Nature Climate Change*, **7 (8)**, 557–562, https://doi.org/10.1038/NCLIMATE3338.

Grams, C. M., L. Magnusson, and E. Madonna, 2018: An atmospheric dynamics perspective on the amplification and propagation of forecast error in numerical weather prediction models: A case study. *Quarterly Journal of the Royal Meteorological Society*, **144 (717)**, 2577–2591, https://doi.org/10.1002/qj.3353.

Hastie, T., R. Tibshirani, and J. Friedman, 2009: *The elements of statistical learning: data mining, inference, and prediction*. 2nd ed., Springer New York, NY, New York, 745 pp., https://doi.org/https://doi.org/10.1007/978-0-387-84858-7.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, **146 (730)**, 1999–2049, https://doi.org/10.1002/qj.3803.

Jolliffe, I. T., 1982: A Note on the Use of Principal Components in Regression. *Applied Statistics*, **31 (3)**, 300, https://doi.org/10.2307/2348005.

Koch, S. E., 1985: Synoptic scale forecast skill and systematic errors in the MASS 2.0 model. *Monthly Weather Review*, **113 (10)**, 1714–1737, https://doi.org/10.1175/1520-0493(1985)113⟨1714:SSFSAS⟩2.0.CO;2.

Lemburg, A., and A. H. Fink, 2022: Identifying Causes of Short-Range Forecast Errors in Maximum Temperature during Recent Central European Heatwaves Using the ECMWF-IFS Ensemble. *Weather and Forecasting*, **37 (10)**, 1885–1902, https://doi.org/10.1175/WAF-D-22-0033.1.

Li, W., B. Pan, J. Xia, and Q. Duan, 2022: Convolutional neural network-based statistical post-processing of ensemble precipitation forecasts. *Journal of Hydrology*, **605 (October 2021)**, 127 301, https://doi.org/10.1016/j.jhydrol.2021.127301, URL https://doi.org/10.1016/j.jhydrol.2021.127301.

Lorenz, E. N., 1956: Empirical Orthogonal Functions and Statistical Weather Prediction. *Technical report Statistical Forecast Project Report 1 Department of Meteorology MIT 49*, **1 (Scientific Report No. 1, Statistical Forecasting Project)**, 52.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus A: Dynamic Meteorology and Oceanography*, **21 (3)**, 289, https://doi.org/10.3402/tellusa.v21i3.10086.

Lorenz, E. N., 1996: Predictablilty: A problem partly solved. *Seminar on predictability*, Vol. 1, 1–18.

Luo, W., Y. Li, R. Urtasun, and R. Zemel, 2016: Understanding the effective receptive field in deep convolutional neural networks. *Advances in Neural Information Processing Systems*, **(Nips)**, 4905–4913.

Messner, J. W., G. J. Mayr, and A. Zeileis, 2017: Nonhomogeneous boosting for predictor selection in ensemble postprocessing. *Monthly Weather Review*, **145 (1)**, 137–147, https://doi.org/10.1175/MWR-D-16-0088.1.

Michel, C., and G. Rivière, 2011: The link between rossby wave breakings and weather regime transitions. *Journal of the Atmospheric Sciences*, **68 (8)**, 1730–1748, https://doi.org/10.1175/2011JAS3635.1.

Michelangeli, P. A., R. Vautard, and B. Legras, 1995: Weather regimes: recurrence and quasi stationarity. *Journal of the Atmospheric Sciences*, **52 (8)**, 1237–1256, https://doi.org/10.1175/1520-0469(1995)052⟨1237:WRRAQS⟩2.0.CO;2.

North, G. R., T. L. Bell, R. F. Cahalan, and F. J. Moeng, 1982: Sampling Errors in the Estimation of Empirical Orthogonal Functions. *Monthly Weather Review*, **110 (7)**, 699–706, https://doi.org/10.1175/1520-0493(1982)110⟨0699:seiteo⟩2.0.co;2.

O'Lenic, E. A., and R. E. Livezey, 1989: Relationships between systematic errors in medium range numerical forecasts and some of the principal modes of low-frequency variability of the Northern Hemisphere 700 mb circulation. *Monthly Weather Review*, **117 (6)**, 1262–1280, https://doi.org/10.1175/1520-0493(1989)117⟨1262:RBSEIM⟩2.0.CO;2.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830, URL http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html%5Cnhttp://arxiv.org/abs/1201.0490.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, **133 (5)**, 1155–1174, https://doi.org/10.1175/MWR2906.1.

Rasp, S., and S. Lerch, 2018a: Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, **146 (11)**, 3885–3900, https://doi.org/10.1175/MWR-D-18-0187.1.

Rasp, S., and S. Lerch, 2018b: Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, **146 (11)**, 3885–3900, https://doi.org/10.1175/MWR-D-18-0187.1.

Rex, D. F., 1950: Blocking Action in the Middle Troposphere and its Effect upon Regional Climate. *Tellus*, **2 (4)**, 275–301, https://doi.org/10.3402/tellusa.v2i4.8603, URL https://doi.org/10.3402/tellusa.v2i4.8603.

Rex, D. F., 1951: The Effect of Atlantic Blocking Action upon European Climate. *Tellus*, **3 (2)**, 100–112, https://doi.org/10.3402/tellusa.v3i2.8617.

Robertson, A. W., and M. Ghil, 1999: Large-scale weather regimes and local climate over the western United States. *Journal of Climate*, **12 (6)**, 1796–1813, https://doi.org/10.1175/1520-0442(1999)012⟨1796:LSWRAL⟩2.0.CO;2.

Rodwell, M. J., D. S. Richardson, D. B. Parsons, and H. Wernli, 2018: Flow-dependent reliability: A path to more skillful ensemble forecasts. *Bulletin of the American Meteorological Society*, **99 (5)**, 1015–1026, https://doi.org/10.1175/BAMS-D-17-0027.1.

Rodwell, M. J., and Coauthors, 2013: Characteristics of occasional poor medium-range weather forecasts for Europe. *Bulletin of the American Meteorological Society*, **94 (9)**, 1393–1405, https://doi.org/10.1175/BAMS-D-12-00099.1.

Storch, H. v., and F. W. Zwiers, 2002: *Statistical Analysis in Climate Research*. Statistical Analysis in Climate Research, https://doi.org/10.1017/cbo9780511612336.

Stoss, L. A., and S. L. Mullen, 1995: The dependence of short-range 500-mb height forecasts on the initial flow regime. *Weather & Forecasting*, **10 (2)**, 353–368, https://doi.org/10.1175/1520-0434(1995)010⟨0353:TDOSRM⟩2.0.CO;2.

Taillardat, M., O. Mestre, M. Zamo, and P. Naveau, 2016: Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review*, **144 (6)**, 2375–2393, https://doi.org/10.1175/MWR-D-15-0260.1.

Tibaldi, S., and F. Molteni, 1990: On the operational predictability of blocking. *Tellus, Series A*, **42 A (3)**, 343–365, https://doi.org/10.3402/tellusa.v42i3.11882.

Van Der Wiel, K., H. C. Bloomfield, R. W. Lee, L. P. Stoop, R. Blackport, J. A. Screen, and F. M. Selten, 2019: The influence of weather regimes on European renewable energy production and demand. *Environmental Research Letters*, **14 (9)**, https://doi.org/10.1088/1748-9326/ab38d3.

Vannitsem, S., and Coauthors, 2021: Statistical Postprocessing for Weather Forecasts. *Bulletin of the American Meteorological Society*, E681–E699.

Veldkamp, S., K. Whan, S. Dirksen, and M. Schmeits, 2021: Statistical postprocessing of wind speed forecasts using convolutional neural networks. *Monthly Weather Review*, **149 (4)**, 1141–1152, https://doi.org/10.1175/MWR-D-20-0219.1.

Wallace, J. M., Yuan Zhang, and Kai-Hon Lau, 1993: Structure and seasonality of interannual and interdecadal variability of the geopotential height and temperature fields in the Northern Hemisphere troposphere. *Journal of Climate*, **6 (11)**, 2063–2082, https://doi.org/10.1175/1520-0442(1993)006⟨2063:SASOIA⟩2.0.CO;2.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*, Vol. 100. Academic Press, https://doi.org/10.2307/2669579, URL https://doi.org/10.1016/C2017-0-03921-6.

Xiang, L., J. Xiang, J. Guan, L. Zhang, Z. Cao, and J. Xia, 2022: Spatiotemporal forecasting model based on hybrid convolution for local weather prediction post-processing. *Frontiers in Earth Science*, **10 (September)**, 1–14, https://doi.org/10.3389/feart.2022.978942.